VERBAL INFORMATION VERIFICATION

Qi Li, Biing-Hwang Juang, Qiru Zhou and Chin-Hui Lee

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, USA
{qli,bhj,qzhou,chl}@research.bell-labs.com

ABSTRACT

Traditionally, speaker authentication has focused on two categories of techniques: speaker verification and speaker identification. In this paper, we introduce a third category called *verbal information verification* (VIV) in which a claimed speaker's utterances are verified against the key information in the speaker's registered profile to decide whether the claimed identity should be accepted or rejected. The proposed VIV technique can be used independently or combined with the traditional speaker verification techniques to achieve flexible and improved speaker authentication. Instead of accomplishing VIV through recognizing the key information, the proposed VIV algorithm is based on the concept of *sequential utterance verification*. In a telephone speaker authentication experiment on 100 speakers and using three pass-utterances in response to three categories of questions, the proposed VIV system achieved 0.00% equal-error rate, compared to 30% false rejection rate on an automatic speech recognition approach.

1. INTRODUCTION

Verbal information verification (VIV) is to verify spoken information against the key information in a given user's profile, such as mother's maiden name, birth place, residence address and so on. Each representing an "information field" in the profile. Verbal information in an utterance is accepted if it contains the correct information according to the target content. One of the important applications of VIV is remote speaker authentication for bank, telephone card, credit card, benefit, and other account accesses. In these applications, a VIV system makes decision on either accepting or rejecting a speaker based on the speaker's spoken personal information. This is similar to current telephone banking procedures: after an account number is provided, an operator verifies a user by asking some personal information, such as birth date, address, home telephone number, etc. A user has to answer the questions correctly in order to gain access to his or her account. Similarly, a dialog VIV system can prompt questions with a text-to-speech synthesizer, and verify spoken information automatically.

A major difference between speaker recognition and VIV in speaker authentication is that speaker recognition inspects speakers' speech characteristics while VIV inspects speakers' verbal content. The difference can be further discussed in three aspects. First, both speaker identification and speaker verification need to train speaker dependent (SD) models or classifiers while VIV can use speaker independent (SI) acoustic models for the purpose of verbal content decoding. Second, speaker recognition needs an enrollment session to record SD speech data, and time to train SD models while VIV does not. The user profiles are created when users' accounts are set up. Third, in speaker verification, the system can reject an imposter who uses a true speaker's spoken password while, in VIV, it is the speakers' responsibility to protect their personal information from impostors. VIV, however, can be used for automatic enrollment of speaker recognition systems, or can be used together with speaker recognition or verification to meet higher security requirements.

2. APPROACHES

We have investigated two kinds of approaches for VIV using the techniques of automatic speech recognition (ASR), Fig. 1, and utterance/speaker verification, respectively.

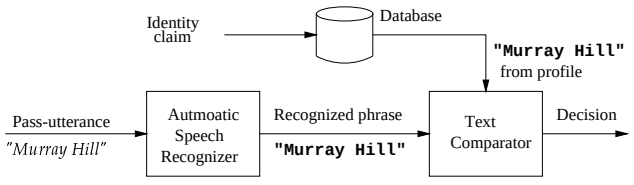

Figure 1: Automatic speech recognition approach to VIV

In an ASR experiment, a VIV system was built for speaker authentication with the three questions asked in a row as shown in Fig. 2. A speaker is accepted if all three questions are answered correctly, or rejected as soon as the recognized result does not match the profile. The recognition rates at each stage are also listed in Fig. 2. For all 100 true speakers, the false rejection rate on all three

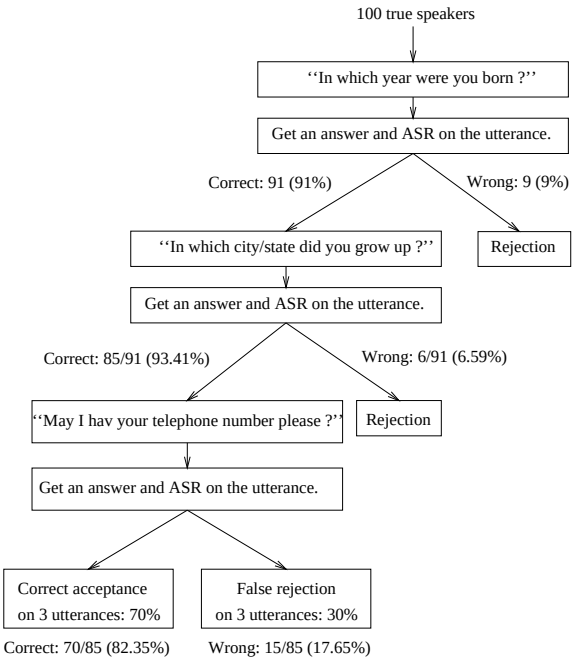

Figure 2: Verbal information verification by ASR. FR = 30% while FA = 0% on three utterances.

questions was 30% with the false acceptance rate on three questions kept at 0%. Three sets of grammar and vocabulary files were used to recognize the pass-utterances respectively. The grammars include the rules for years, city/state names, and 10-digit telephone numbers.

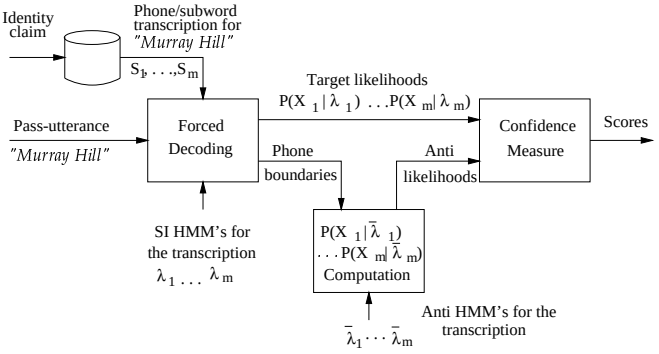

Figure 3: Verification approach to VIV

From the verification point of view, the above ASR approach does not effectively utilize the information in the profile. As shown in Fig. 3, we can use the subword transcription of the text in a profile (a known correct answer) to decode an utterance, i.e. so called forced decoding. This will give us the subword segmentation boundaries. Then, hypothesis test techniques can be applied to decide either to accept or reject an utterance. This approach is similar to the verification techniques used in speaker verification [1, 2] and utterance verification [3, 4, 5, 6, 7]. The rest of this paper will focus on the verification approach.

2.1. Normalized Confidence Measure

During the hypothesis test for segmented subwords, confidence scores are calculated for decision. Several confidence measures have been used in utterance verification [6, 8]. We propose a *normalized confidence measure* for some practical reasons which will be discussed below. Following the concept of *inspection by variable* [9] in hypothesis test, we define a confidence measure for a decoded subword n in an observed speech segment O_n as

$$C_n = \frac{\log P(O_n|\lambda_n^t) - \log P(O_n|\lambda_n^a)}{\log P(O_n|\lambda_n^t)}, \qquad (1)$$

where λ_n^t and λ_n^a are the corresponding target and anti models for subword unit n respectively, $P(.)$ is the likelihood, and assume $\log P(O_n|\lambda_n^t) > 0$. The *normalized confidence measure* for an utterance with N subwords is

$$M = \frac{1}{N}\sum_{n=1}^{N} f(C_n), \qquad (2)$$

where

$$f(C_n) = \begin{cases} 1, & \text{if } C_n \geq \theta; \\ 0, & \text{otherwise,} \end{cases} \qquad (3)$$

and θ is a common subword threshold for all subwords. M, $0 \leq M \leq 1$, can be interpreted as a percentage of acceptable subwords in an utterance, e.g. $M = 0.8$ means that 80% of the subwords in an utterance are acceptable. Thus, an utterance threshold can be determined based on the specifications of system performance and robustness. Two properties of the normalized measure may be desirable: a common subword threshold, and the physical meaning on the utterance threshold which can even be determined by users or customers. We note that the confidence measure is specially defined for VIV.

2.2. Multiple Thresholds and Tolerance Intervals

Once an utterance score is determined, a decision can be made to either reject or accept an utterance, i.e.

$$\begin{cases} \text{Acceptance:} & M_i \geq T_i; \\ \text{Rejection:} & M_i < T_i, \end{cases} \qquad (4)$$

where M_i and T_i are the corresponding confidence score and threshold for utterance i. For a multiple-utterance VIV system, either one global threshold, i.e. $T = T_1 \ldots = T_i$, or multiple thresholds, i.e. $T_1 \neq T_2 \ldots \neq T_i$, can be used. The thresholds can be either context (i.e. information field) dependent (CD) or context independent (CI). It can also be either speaker dependent (SD) or speaker independent (SI).

For robust verification, we define the logic of using two global thresholds for a multiple-question trial as follows.

$$T_i = \begin{cases} T_{\text{low}}, & \text{when } T_{\text{low}} \le M_i < T_{\text{high}} \text{ at the first time} \\ & \text{and } T_{\text{low}} \text{ can be used only once,} \\ T_{\text{high}}, & \text{otherwise,} \end{cases} \tag{5}$$

where T_{low} and T_{high} are two thresholds. Eq. (5) means T_{low} can be used only once in one verification trial. Thus, if a speaker has only one lower score in answer utterances, the speaker still has the chance to pass the overall verification trial. This is useful in noisy environments or for speakers who may not speak consistently.

To further improve the performance of a VIV system, we use speaker and context dependent thresholds. To guarantee no false rejection, the upper bound of the threshold for utterance i of a speaker can be selected as

$$t_i = \min\{M_{i,j}\}, \qquad j = 1, ..., J, \tag{6}$$

where $M_{i,j}$ is the confidence score for utterance i on the jth trials. Due to the changes on voice, channels, and environment, the same speaker may have different scores even for the same context utterance. We define an *utterance tolerance interval* τ as

$$T_i = t_i - \tau, \tag{7}$$

where t_i is defined as in Eq. (6), $0 \le \tau < t_i$, and T_i is a CD utterance threshold for Eq. (4). By applying the tolerance interval, a system can still accept a speaker although his or her utterance score M_i on the same context is lower than before. For example, a speaker's minimal confidence measure on the answer to the ith question is $t_i = 0.9$. If a VIV system is designed with $\tau = 0.06\%$, we have $T_i = 0.9 - 0.06 = 0.84$. This means that the speaker still can be accepted as long as 84% of the subwords of utterance i are acceptable.

In the system evaluation, τ can be reported with error rates as a guaranteed performance interval. On the other hand, in the system design, τ can be used to determine the thresholds based on system specifications. For example, a bank authentication system may need a smaller τ to ensure lower false acceptance rates at a higher security level while a voice mail system may select a larger τ to reduce false rejection rates for a user friendly security access.

2.3. Error Rates on Sequential Utterance Test

As is well known, when performing a test on a single utterance, one may commit one of two errors: rejecting the hypothesis when it is true – false rejection, or accepting it when it is false – false acceptance. When more than one utterance are given sequentially for speaker authentication test, we define *false rejection on K utterances* ($K \ge 1$) to be the case where a true speaker is rejected in any one of the K hypotheses tests, and *false acceptance on K utterances* to be the case where an imposter is accepted through all K hypothesis tests. An *equal-error rate on K utterances* (EER) is the rate on which false rejection and false acceptance on K utterances are equal. For the following experiments, we set $K = 3$.

3. DATABASES AND HMM'S

The experimental database includes 100 English speakers. Each speaker has 3 utterances as the answers to three questions: "In which year were you born?", "In which city and state did you grow up?", and "May I have your telephone number please?" It is a biased database since 26% of the speakers are with birth years in the 1950's, and 24% are in the 1960's. We note that there is only one digit different among those birth years. In the city and state names, 39% are "..., New Jersey", and 5% of the speakers use exactly the same address "Murray Hill, New Jersey", which means the verification system will give the same results on these addresses. Thirty eight percent (38%) of telephone numbers start from "908 582 ...", which means that at least 60% of the context of the telephone numbers are exactly the same for those numbers. Also, some of the speakers have accents, and some cities and states are in foreign countries.

In our speaker authentication experiments, a speaker is considered as a true speaker when the speaker's utterances are verified against his or her profile. On the other hand, the speaker is considered as an impostor when the utterances are verified against other speakers' profiles. Thus, for each speaker, we have three utterances from the true speaker and 99×3 utterances from the impostors.

We used a set of 1117 right CD HMM's as target models, and a set of 41 CI anti-phone HMM's as anti models [3, 6, 10]. The CD models were trained by a discriminative learning algorithm for minimal error classification [10, 11]. ASR and forced decoding were done by the target models. Both the target and anti models were used for verification.

4. EXPERIMENTAL RESULTS

When one common SI utterance threshold was applied to all speakers and all questions, the EER was less than 1%. When the two thresholds as in Eq. (5) were applied, the EER was 0.57%.

To further improve the performance of VIV, we applied SD and CD thresholds defined in Eqs. (6) and (7) to this experiment. There were three thresholds associated with the three questions for each speaker. The thresholds were determined by first finding the t_i's as in Eq. (6) to guarantee 0% false rejection rates. Then, the thresholds were shifted to find the false acceptance rates on different tolerance intervals τ as defined in Eq. (7). The relation between tolerance values

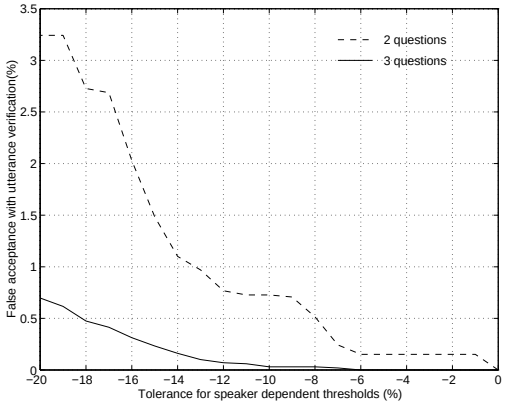

Figure 4: The system performances when FR = 0%.

and false acceptance rates on the three questions using the normalized confidence measure are shown in Fig. 4. The two curves represent the performances of the VIV systems using a set of two and three questions for speaker authentication while maintaining the false rejection rate to be 0.00%. The decision logic is that whenever one answered utterance is not acceptable, the corresponding speaker is rejected and no further utterances will be evaluated.

As shown in Fig. 4, using two questions, we can have a 0% EER only when the tolerance interval τ is 0, which means that when a true speaker's utterance score as in Eq. (2) is lower than before the speaker will be rejected; with three questions, the VIV system gave 0.00% EER with 6% tolerance interval, which means when a true speaker's utterance scores are 6% lower than before (or unacceptable phones are 6% more than before), the speaker can still be accepted while all impostors in the database can still be rejected correctly. This tolerance interval gives the room for variation in the true speaker's score to ensure a reliable performance.

5. DISCUSSIONS AND CONCLUSIONS

In real speaker authentication applications, to avoid impostors using a speaker's personal information which is just uttered, a VIV system may randomly ask a subset of personal information for each access. For example, the users are registered 6 items, and each time the system randomly picks 3 to verify. Or, the system may ask some dynamic information recorded from the past transactions, such as the date or the amount of the last deposit.

A practical VIV system may apply SI thresholds (5) to new users and switch to SD thresholds when the thresholds in (6) are determined. Such SD thresholds can be stored in credit cards or phone cards for user authentication applications.

In conclusions, verbal information verification opens a new area for speaker authentication. A normalized confidence measure and associated hypothesis tests were presented in this paper. In the speaker authentication experiments with three questions prompted sequentially, the proposed VIV system achieved a 0.00% equal-error rate plus a 6% tolerance interval.

6. REFERENCES

[1] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Proc. IEEE ICASSP*, May 1996.

[2] Q. Li, S. Parthasarathy, and A. E. Rosenberg, "A fast algorithm for stochastic matching with application to robust speaker verification," in *Proc. IEEE ICASSP*, April 1997.

[3] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. Speech and Audio Process.*, vol. 4, Nov. 1996.

[4] R. A. Sukkar, A. R. Setlur, M. G. Rahim, and C.-H. Lee, "Utterance verification of keyword string using word-based minimum verification error (WB-MVE) training," in *Proc. IEEE ICASSP*, May 1996.

[5] A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting recognition errors via discriminative utterance verification," in *Proc. ICSLP*, Oct. 1996.

[6] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Combining key-phrase detection and subword-based verification for flexible speech understanding," in *Proc. IEEE ICASSP*, May 1997.

[7] M. G. Rahim, C.-H. Lee, and B.-H. Juang, "Robust utterance verification for connected digits recognition," in *Proc. IEEE ICASSP*, May 1995.

[8] E. Lleida and R. C. Rose, "Efficient decoding and training procedures for utterance verification in continuous speech recognition," in *Proc. IEEE ICASSP* May 1996.

[9] E. L. Lehmann, "Testing statistical hypotheses," NY:John Wiley & Son, 1959.

[10] C.-H. Lee, B.-H. Juang, W. Chou and J. J. Molina-Perez, "A Study on Task-Independent Subword Selection and Modeling for Speech Recognition," *Proc. ICSLP*, Oct. 1996.

[11] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Process.*, vol. 5, May 1997.